Cell Cycle 11:4, 658-667; February 15, 2012; © 2012 Landes Bioscience

# New insights into replication origin characteristics in metazoans

Christelle Cayrou,[1,†] Philippe Coulombe,[1,†] Aurore Puy,[1] Stéphanie Rialle,[1] Noam Kaplan,[2] Eran Segal[2] and Marcel Méchali[1,*]

[1]Institute of Human Genetics; CNRS; Montpellier, France; [2]Department of Computer Science and Applied Mathematics; Weizmann Institute of Science; Rehovot, Israel

[†]These authors contributed equally to this work.

**We recently reported the identification and characterization of DNA replication origins (Oris) in metazoan cell lines. Here, we describe additional bioinformatic analyses showing that the previously identified GC-rich sequence elements form origin G-rich repeated elements (OGREs) that are present in 67% to 90% of the DNA replication origins from Drosophila to human cells, respectively. Our analyses also show that initiation of DNA synthesis takes place precisely at 160 bp (Drosophila) and 280 bp (mouse) from the OGRE. We also found that in most CpG islands, an OGRE is positioned in opposite orientation on each of the two DNA strands and detected two sites of initiation of DNA synthesis upstream or downstream of each OGRE. Conversely, Oris not associated with CpG islands have a single initiation site. OGRE density along chromosomes correlated with previously published replication timing data. Ori sequences centered on the OGRE are also predicted to have high intrinsic nucleosome occupancy. Finally, OGREs predict G-quadruplex structures at Oris that might be structural elements controlling the choice or activation of replication origins.**

**Key words:** DNA replication origins, DNA synthesis, G-quadruplex, nucleosome, CpG islands, transcription

Submitted: 12/15/11

Accepted: 12/16/11

http://dx.doi.org/10.4161/cc.11.4.19097

*Correspondence to: Marcel Méchali; Email: mechali@igh.cnrs.fr

## Introduction

DNA replication initiates at discrete sites called replication origins (Oris), which should be activated only once at each cell cycle to avoid amplification and maintain genome integrity. In bacteria, yeast and viruses, the structure and regulation of Oris are rather well understood and are characterized by specific DNA sequence elements. In Metazoa, Oris occur at specific locations,[1,2] but their genetic characteristics remain unclear.

In bacteria, control of replication initiates by the binding of DnaA, the initiator of chromosome replication, to the DnaA boxes, which form clusters of three or more elements and are located mainly in the *E. coli* Ori. DnaA forms a nucleoprotein complex with Ori that results in the loading of different components of the replisome, leading to DNA replication.[3] The archaeal DNA replication machinery has many similarities and is a simplified form of the one in eukaryotes.[4] The archaeal chromosome generally contains one Ori that has autonomously replicating sequence activity, although some Achaea contain multiple Oris.[5] These Oris (termed ORBs) are well conserved across many archaeal species and present inverted repeat sequence elements that are bound by a complex that is similar to the eukaryotic origin recognition complex (ORC).[6] ORB is a ~36 bp CG-rich sequence near an AT-rich region that might act as a melting-prone site. From viruses to budding yeast, a specific sequence that is necessary for Ori activity has been identified. However, the features of this sequence vary from one organism to the other. DNA viruses are the smallest self-replicating entities, and specific palindromic sequences are required for their replication.[7,8] In most cases, the main initiating protein is encoded by the viral genome itself. Its binding to the Ori is also actively involved in the regulation of viral transcription and chromosomal segregation. Notably, the involvement of

transcription factors in DNA replication has been clearly demonstrated in adenoviruses, papovaviruses, including the simian virus 40 (SV40), and papillomaviruses.[9,10]

In budding yeast, replicators and origins are together defined as ARSs (autonomously replicating sequences), upon which multiple initiation proteins are assembled stepwise. This is the only eukaryote in which ORC specifically recognizes a 17 bp T-rich consensus sequence called ACS (ARS consensus sequence).[11] Although this motif is necessary for DNA replication, it is not sufficient for Ori function.[12] In contrast to *S. cerevisiae*, *S. pombe* Oris do not contain a core consensus sequence essential for their function.[13] However, *S. pombe* Oris are AT-rich (from 0.5 to 3 kb in length) and contain several functionally important DNA sequence elements for their activity.[14] In a genomic context, a 30 bp-long poly-A/T track appears sufficient to specify replication initiation.[15]

Until recently, how Oris are defined in metazoans remained elusive despite considerable efforts to unravel a replication origin code. The structure and initiator role of ORC is conserved in all eukaryotes, and in vitro and in vivo studies indicate that although ORC is essential for DNA replication, it does not show noticeable DNA sequence specificity in vitro.[16] It seems, however, that the pre-replication complex (including ORC and other proteins of the replication initiation) may have a greater affinity for specific regions within Oris.[16] Oris appear to have variable features, since they can be extremely site-specific, like the human Lamin B2 or c-Myc Oris,[17,18] or have a broad site specification, like the DHFR Ori.[19] We recently characterized up to 2,412 Oris on chromosome 11 in mouse ES cells and 6,184 Oris in the Drosophila genome by Nascent Strands (NS) purification and mapping by microarrays.[2] Our sequences analysis unexpectedly revealed specific G-rich motifs in both mouse and Drosophila Oris near initiation sites.[2] Here, we describe new bioinformatic analyses, showing that these origin G-rich repeated elements (OGREs) allow a good prediction of Oris both in the mouse and Drosophila genome. OGREs are also present in the majority of the previously characterized Oris in human cells.[1] Interestingly, initiation of DNA synthesis occurs preferentially at the 3' of OGREs. These data give a strong support to the hypothesis that sequence-specific elements are also involved in Ori recognition or function in metazoans.

Many G-rich motives are predicted to form quadruplexes (G4). G4 formation involve the association of four guanines into a cyclic Hoogsteen hydrogen bonding arrangement in which each guanine shares a hydrogen bond with its neighbor,[20] and it can be predicted by computational techniques from DNA sequences.[21] Predicted G4 are not randomly located in genomes and play important physiological roles. For instance, they are found in telomeres[22] and have been implicated in regulating transcription, translation and replication.[23-26] We found that G4 are highly associated with Oris. These data suggest that these motifs have a structural role in Ori localization or activation.

## Results

**Metazoan replication origins are characterized by OGRE.** We recently mapped active Oris at the genomic scale in both Drosophila and mouse cells.[2] Several thousand novel Oris were identified, thus allowing the characterization of their general features. We reported that Oris were significantly enriched in genes vs. intergenic regions and also found a large body of Oris between genes. Specifically, Oris tended to be more abundant in actively transcribed genes. We also noticed a strong Ori enrichment at transcription start sites (TSS) in all mouse cell lines analyzed but not in Drosophila Kc cells. However, the link with transcription might be indirect, as the CpG islands (CGI) embedded in TSS, rather than the TSS on their own, seemed to be important. Indeed, CGIs outside TSS can also be Oris, whereas TSS without CGI are not.

We also showed that Oris contain specific nucleotide sequences. Indeed, consensus motifs were associated with both Drosophila and mouse Oris. As these motifs are G-rich sequences and highly associated with Oris, we propose to name them origin G-rich repeated elements (OGREs, **Fig. 1A**).

To further characterize the relationship between OGREs and Oris, we evaluated their genomic association in the mouse and Drosophila genome using the FIMO software (see Materials and Methods). OGRE occurrences were strongly and significantly associated with metazoan Oris (**Fig. 1B and C**). Specifically, 80–90% of mouse Oris possessed at least one OGRE. Drosophila Oris also exhibited a significant association with OGREs. Moreover, mouse and fly OGREs were largely interchangeable, suggesting that a common conserved mechanism acts to specify active Oris in metazoans. Several hundred new Oris were recently mapped on the human ENCODE regions,[1] which correspond to about 1% of the human genome. We observed that the OGRE occurences were also present in the vast majority of these Oris (**Fig. 1D**).

Finally, we determined the density of both Ori and OGRE occurrence in mouse chromosome 11 and found that they were strongly correlated (**Fig. 1E**). This suggests that the OGRE motif might play a role in specifying Oris in large-scale domains. Moreover, OGRE density correlated also with replication timing,[27] as higher OGRE density tended to be associated with early replicating domains.

**Initiation of replication takes place at a specific position downstream of the OGRE motif.** We then evaluated the NS signal profile around all OGRE motifs that occur in the mouse and Drosophila genomes (**Fig. 2A and B**) and found a strong initiation peak located ~280 bp (mouse) and 160 bp (Drosophila) downstream of the OGRE occurrences.

Recently, we reported the presence of a nucleotide skew or bias around NS peaks[2] that was deduced by analyzing the sequence of only one DNA strand. Specifically, Oris show an over-representation of G nucleotides upstream and of C nucleotides downstream of the NS peaks. Drosophila Oris show a T/A skew in addition to the G/C skew. We thus investigated whether such nucleotide skew was a unique characteristic of Oris by calculating all the regions exhibiting a nucleotide skew in the mouse and Drosophila genomes (see Materials and Methods). This was not the case, as the nucleotide skew was present not only in Oris, as previously reported,

but also at other areas, indicating that the nucleotide skew alone is not a predictive value for Oris. However, the OGRE motif on its own had a strong skew, and its presence near NS peaks could be responsible for the observed skew at Oris. To test this hypothesis directly, we oriented or not Oris relative to the sense of the OGRE motifs. Non-oriented Oris showed the characteristic G/C skew around the NS peak (**Fig. 2C**). However, when oriented relative to the OGREs, Oris had only a strong G-skew at the 5' of the peak without the C-skew at the 3' (**Fig. 2D**). This clearly indicates that the nucleotide skew observed near NS peaks is due to the presence of OGREs. In agreement with this hypothesis, the nucleotide skew was located ~280 pb 5' of the NS peak in ES cells, as reported in **Figure 2A**.

This analysis strongly suggests that the OGRE motif might be a genetic element that drives the downstream initiation of DNA synthesis (**Fig. 2E**), possibly by assembling a still-unknown complex that permits the loading of DNA polymerases and the activation of DNA synthesis.

**Bimodal Oris found at CGI are associated with OGREs occurring on both strands.** In our previous work, we also identified a strong association between Oris and CGI in mice and CGI-like in Drosophila. Indeed, ~60% of CGI correlated with active Oris.[2] In comparison, OGREs had a lesser predictive power (41%, **Fig. 3A**, left diagram). However, when combining CGI and OGRE occurrence, we could predict a substantial repertoire of Oris (~25%, data not shown) with high confidence. Indeed, about 90% of the OGRE/CGI occurrences were associated with Oris (**Fig. 3A**, center).

Moreover, as CGI-associated Oris display a bimodal NS profile, suggesting the presence of two replication initiation sites,[2] we asked whether the link between OGRE and CGI could explain this bimodality. The NS signal profiles around the CGI-positive and -negative OGRE occurrences were computed. Strikingly, OGRE associated with CGI displayed a bimodal NS pattern, whereas OGRE not associated with CGI had a single 3' peak (**Fig. 3B**). The predominant occurrence of the OGRE motif on one strand in unimodal (CGI-negative) Oris and on both strands in bimodal (CGI-positive) Oris could explain this result (**Fig. 3D**). Indeed, we observed that the vast majority of CGI-positive Oris contained OGREs positioned on the (+) strand and also on the (-) strand (**Fig. 3C**). This was not the case for Oris that were not associated with CGI.

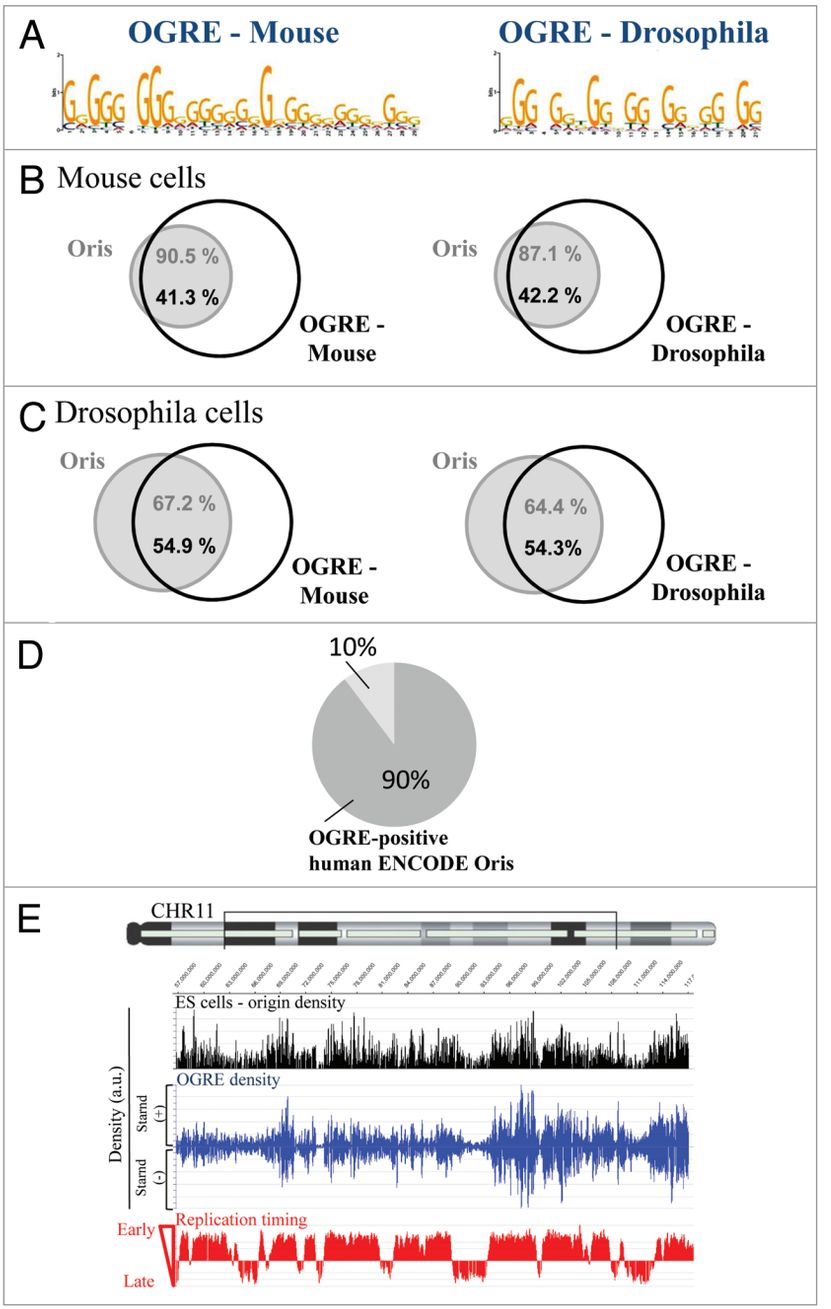


**Figure 1.** Metazoan replication origins (Oris) contain an Origin G-rich repeated element (OGRE). (A) Description of the consensus elements (OGRE) found in Oris from mouse ES (left part) and Drosophila Kc cells (right part). Venn diagrams illustrate the association between Oris in mouse ES (B) and Drosphila Kc cells (C) and OGRE occurrences. (D) Presence of the mouse OGRE motif in Oris mapped in the ENCODE regions of HeLa cells. The proportion of OGRE-positive and -negative Oris is indicated. (E) OGRE density correlates with Oris density and early replication timing.

These data indicate that CGI contain multiple OGREs in different orientations that could force two potential sites of initiation of replication.

**OGREs are predicted to have high intrinsic nucleosome occupancy.** An interesting concept put forward in recent

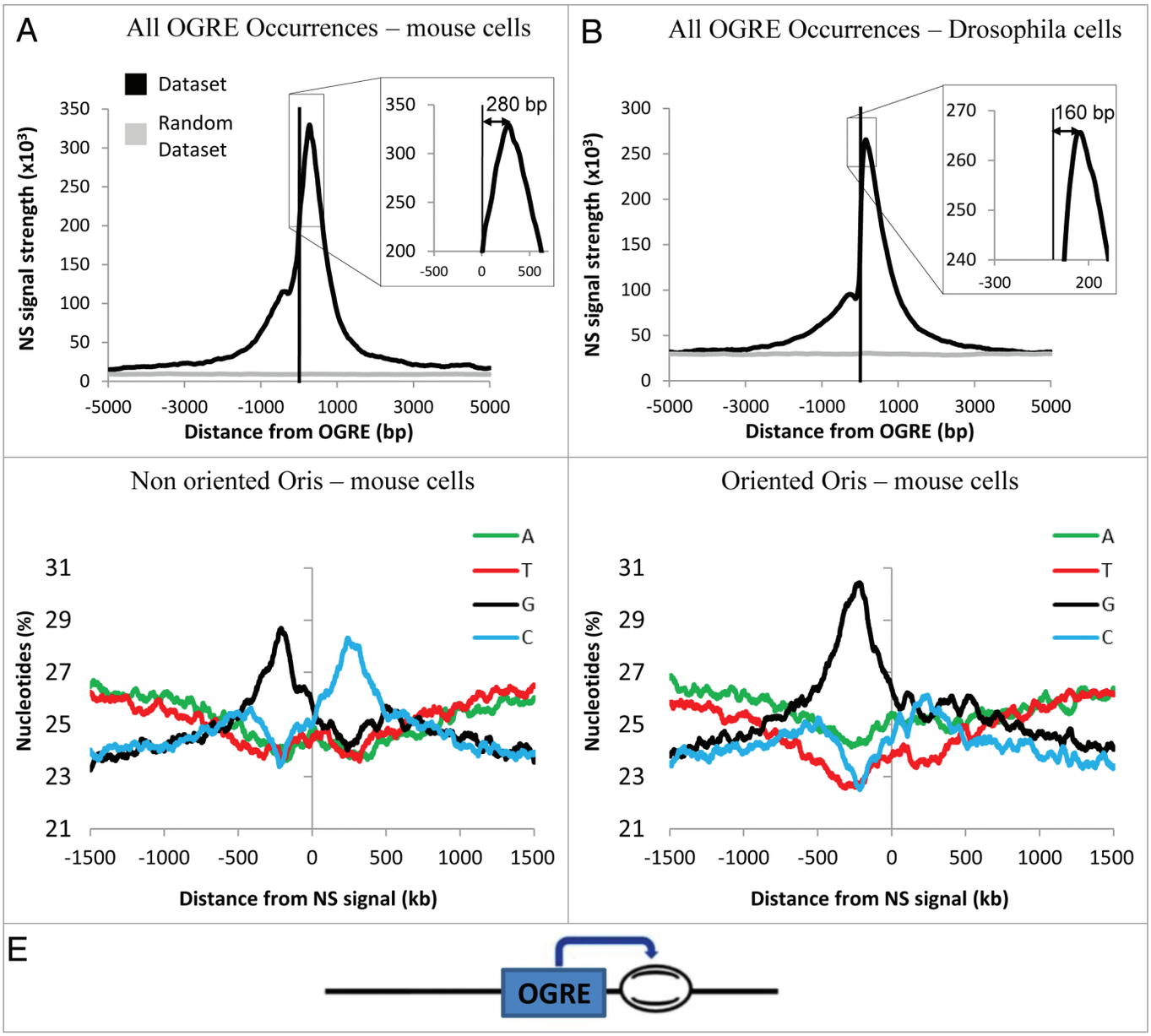


**Figure 2.** OGRE is localized upstream of metazoan Ori peaks. Nascent Strands (NS) enrichment at OGRE is shown for mouse (A) and Drosophila (B) cells. A strong NS peak is found ~280 and ~160 nucleotides 3' of the overall mouse or Drosophila OGRE occurrences, respectively. NS signals were not associated with randomized OGRE occurrences. The enrichment value is the negative log of the combined p-value associated with the NS signal. (C) OGRE-positive mouse Oris were aligned and centered on the NS peaks. The nucleotide distribution was calculated ±1 kb of the NS peak. Note the presence of G-(black) and C-rich (blue) sequences 5' and 3' of the NS peaks. In this situation, Oris were not oriented relative to the OGREs. (D) When mouse Oris were oriented relative to the OGREs, the nucleotide distribution around NS peaks was characterized by the presence of a strong G-rich strand (black) 5' of the NS peak. (E) Schematic representation of OGRE organization relative to the NS peak.

years is the idea that the DNA sequence itself can dictate nucleosome occupancy to some extent.[28] Models based on this concept allowed predicting, for instance, the global nucleosome occupancy at *S. cerevisiae* Oris, with results that were in agreement with the in vivo data,[11] although occupancy measured in vivo is also influenced by other mechanisms.[29] Due to the lack of in vivo measurements, we decided to evaluate the average nucleosome occupancy of OGREs occurences at Oris using a computational model of nucleosome sequence preferences.[28] We found that Ori sequences centered on the OGREs were predicted, on average, to have high intrinsic nucleosome occupancy in both mouse (**Fig. 4A**) and Drosophila (**Fig. 4B**) genomes using both mouse and Drosophila OGRE motifs (data not shown). The predicted nucleosome occupancy was also higher than average, ~200 bp both up and downstream of the OGRE motifs.

**OGREs are linked to structural G-quadruplex (G4) elements.** G-rich sequences have the potential to form G-quadruplexes (G4),[20] which are

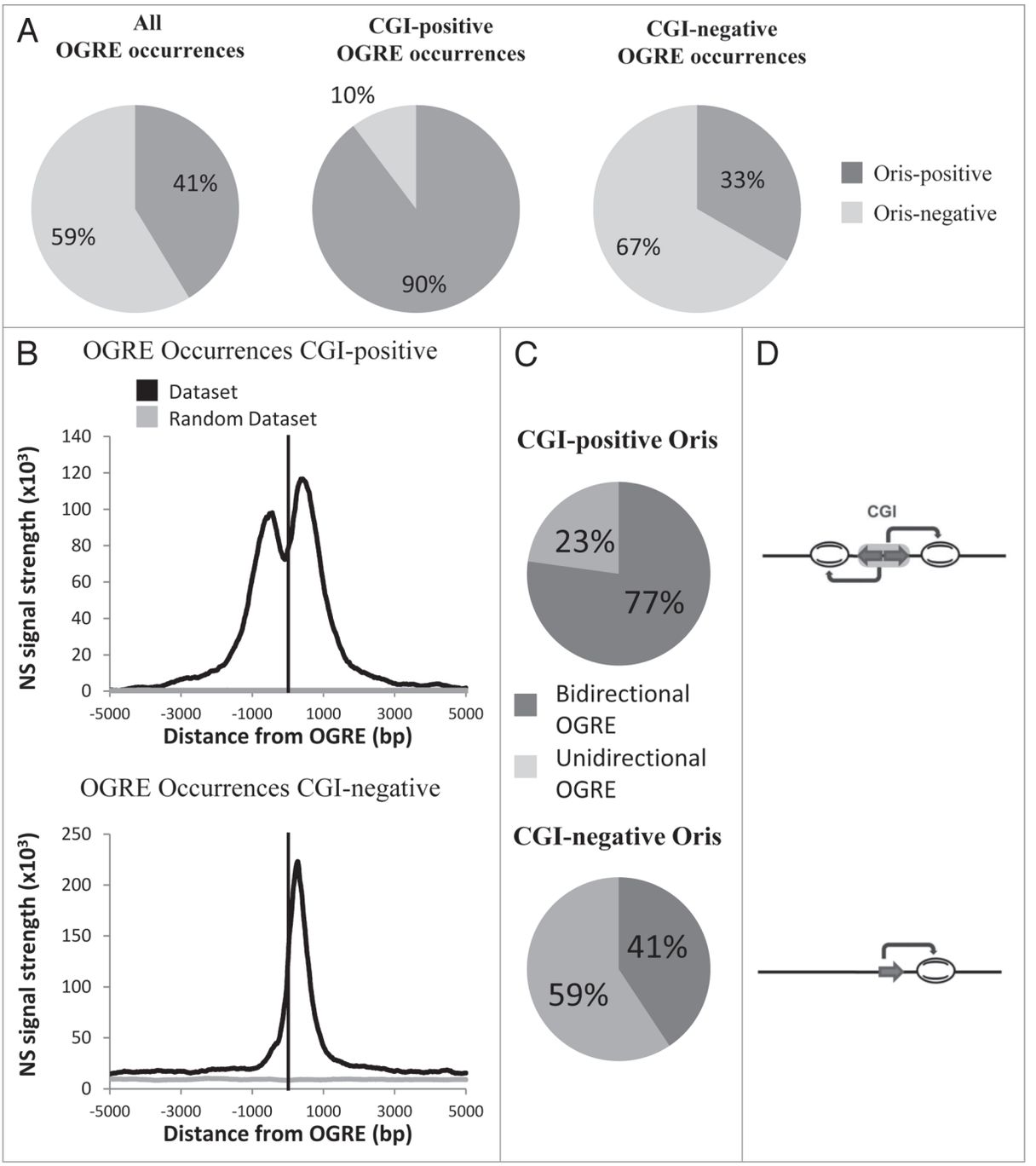


**Figure 3.** CpG Islands (CGI) and OGRE predict metazoan Oris. (A) Association of mouse ES cell Oris with all OGRE occurrences (left part) and with CGI-positive (middle part) or CGI-negative (right part) OGRE occurrences. (B) The Nascent Strands (NS) enrichment at CGI-positive (upper part) and CGI-negative (lower part) OGRE occurrences is shown. Note that the dual peak is only seen in the CGI-positive OGRE occurrences. (C) The presence of OGRE occurrences on both strands (bidirectional OGRE) and on one strand (unidirectional OGRE) was analyzed for CGI-positive (upper part) and CGI-negative (lower part) Oris. Note that CGI-positive Oris tend to have more OGRE occurrences on both strands (upper part), whereas CGI-negative Oris usually have occurrences only on one strand (lower part). (D) Schematic representation of OGRE organization relative to the NS peaks.

single-strand DNA molecules folded in a four-stranded structure connected by small loops. We thus predicted the G4 genomic occurrence using Quadparser[21] and the consensus sequence: $G3N_{1-7}G3N_{1-7}G3N_{1-7}G3$ (**Fig. 5A**). We found that most G4 elements (82%) were predicted to occur at the site of OGREs (**Fig. 5B**), and that 50% of OGRE elements had the potential to form G4 structures. An example of G4-forming OGRE is depicted in **Figure 5C**.

Interestingly, most mouse Oris (1784) were associated with G4 (**Fig. 5D**). Oris with G4 elements also possessed OGREs

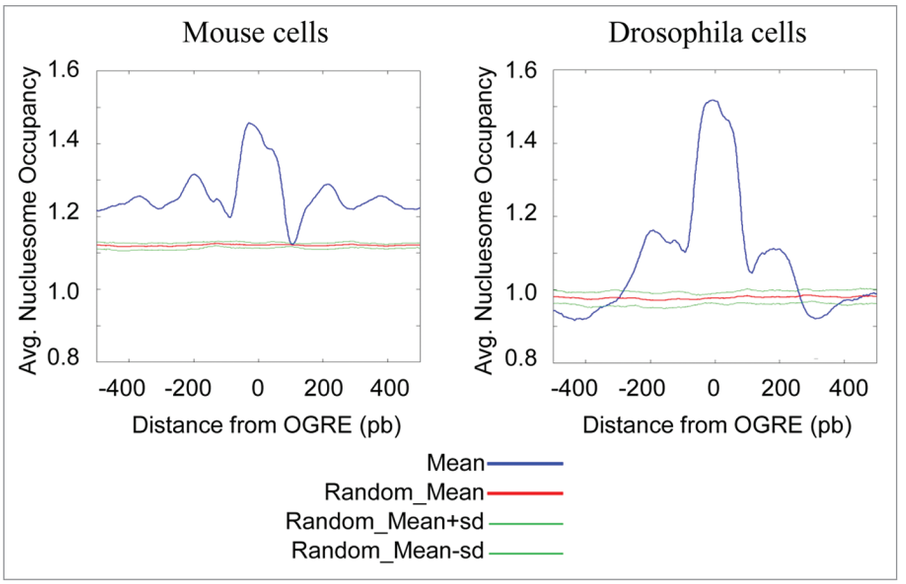


**Figure 4.** OGREs are predicted to have high intrinsic nucleosome occupancy. The average predicted intrinsic nucleosome occupancy was computed for a 1 kb around the most significant OGRE occurrences associated with Oris. Shown is the predicted average occupancy for mouse (A) and Drosophila (B) Oris centered on the mouse OGRE motif. As a control, a similar number of sites were distributed randomly and the average nucleosome occupancy was calculated around these sites. The control randomization was repeated 20 times and its mean and standard deviation are plotted in red and green, respectively. Nucleosome occupancy was normalized to the mean genome occupancy that is 1.

(**Fig. 5E**). Only 1% of Oris contained G4 without OGRE. Importantly, the OGRE-G4 association predicted over 70% of Oris, with a predictability of 51%. Similar to what observed for the OGRE motifs, initiation of DNA synthesis showed a clear peak at ~280 bp downstream the G4 elements (**Fig. 5F**).

These data suggest that OGREs could have some structural features (similar to the G-quadruplex structure) that can favor their function.

## Discussion

We found that 90% of Oris in mouse and human cells and nearly 70% in Drosophila cells possess a specific G-rich sequence called OGRE.[2] This is the first study describing a sequence element that is conserved in different metazoans species. In general, each Ori contains several copies of this motif (only 5% of Oris have a single OGRE), similar to what was observed in other organisms, like Archea, where Oris typically contain multiple ORB repeats. In addition, and in agreement with our data, ORB can be in similar or reversed orientations without affecting binding capacity for the ORC1 dimer.[30] The OGRE motif is probably not a direct binding site for ORC, as our data show instead a depletion of binding sites for mouse ORC2 in this sequence (data not shown). As the replication initiation site occurs at more than 160 bp (Drosophila cells) and 280 bp (mouse cells) downstream of the OGRE, it is possible that a large complex assembles between the OGRE and the ORC binding site, covering this DNA space and governing the entry of DNA polymerases (**Fig. 2C**).

In many organisms, following the assembly of ORC complexes on Oris, the protein core is wrapped by DNA. In *A. pernix*, this conformational change might increase the accessibility of A + T-rich regions, resulting in DNA unwinding and beginning of replication.[30] Atomic force microscopy studies of the *S. pombe* ORC have suggested that approximately 150 bp of DNA are wrapped around the protein complex.[31] DNA wrapping of Ori binding proteins has been observed also with the Drosphila DmORC,[32] and this could explain the distance of 160 bp that we find between the OGRE motif and the NS peak in Kc cells.

By its wrapping around DNA, ORC enhances DNA distortion on one side but with reduced accessibility on the other, thus promoting DNA unwinding on the right-hand side.[33] A study in Drosophila showed that DmORC binds to negatively supercoiled DNA ~30-fold more avidly than to linear or relaxed DNA.[30] This particular distortion of DNA also facilitates the formation of G-quadruplexes.[34] Using bioinformatic prediction tools, we identified a propensity of OGRE sequences to form G-quadruplex structures, which were predicted to occur in more than 70% of mouse Oris. The observation that most metazoan Oris contain motifs with a preferential orientation, as seen in **Figure 2A and B** (except for Oris associated with CGI sequences), consolidates this hypothesis. G-quadruplexes might function as switches, regulating initiation of DNA replication. Their stabilization by specific proteins might maintain the distortion of the DNA complex that stimulates or inhibits the recruiting of the ORC complex to Oris.

We then show that Ori sequences centered on the OGREs are predicted to have high intrinsic nucleosome occupancy. However, Oris in yeast are preferentially located in nucleosome-free region.[11] This contradiction can be explained by the difference in nucleosome occupancy in yeast and human regulatory elements.[35] Notably, many human transcription factor (TF) binding sites have high intrinsic nucleosome occupancy, in contrast to yeast sites, yet both are depleted in vivo. In addition, nucleosomes bind preferentially to GC-rich DNA in vitro,[36] but CGI have lower than average in vivo nucleosome occupancy.[35] Nevertheless, some TF binding sites are GC-rich and have high intrinsic and in vivo nucleosome occupancy, thus arguing for an in vivo preference for nucleosome occupancy at OGRE sequences. Nucleosomes are expected to hinder DNA melting, which is required for initiation of DNA replication, and the NS peak, located ~160 to 300 pb downstream of the OGRE, is a region where intrinsic nucleosome occupancy is predicted to be less abundant than in the OGRE region. Further investigations are

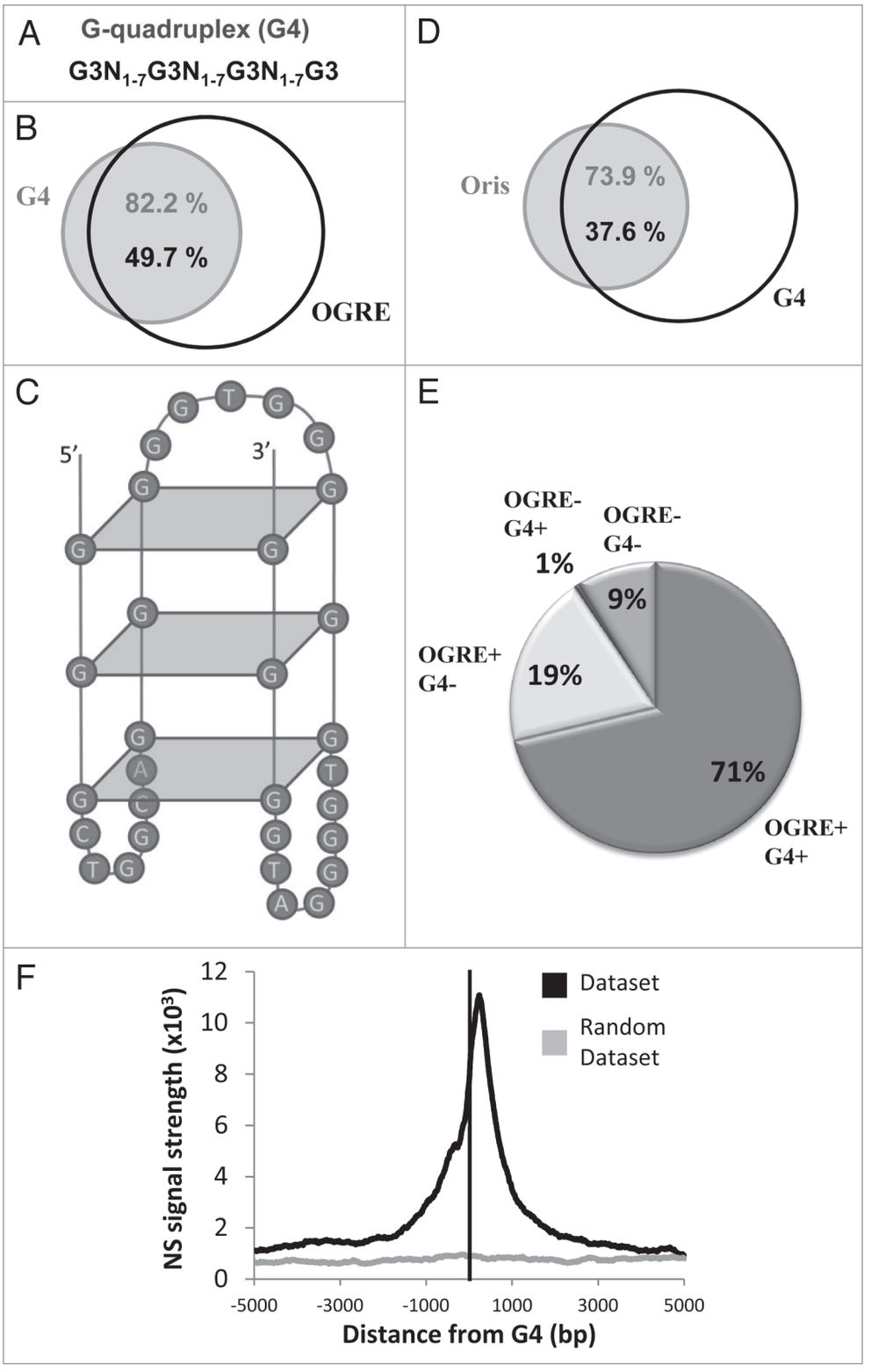


**Figure 5.** G-quadruplexes are specifically found in the majority of metazoan Oris, in association with the OGRE motif. (A) Definition of the G-quadruplexes (G4) that were used in this study. (B) Venn diagram showing the association of OGRE occurrences and G4 in mouse chromosome 11. (C) Predicted G4 formed by an Oris-associated OGRE occurrence. (D) The association between Oris and G4 are illustrated with a Venn diagram. (E) Repartition of Oris from mouse ES cells based on to the presence of OGREs and/or G4. (F) The nascent Strands (NS) enrichment around G4 occurrences is shown. A strong NS peak is found ~260 nucleotides downstream of the G4. A small NS peak is also observed upstream similarly to what described for OGREs (see Fig. 2A and B).

required to evaluate the relevance of the high intrinsic nucleosome occupancy at OGREs and the nucleosome positioning at Oris around OGREs.

Several proteins of the large helicases family can unwind G-quadruplexes. There is now good evidence, particularly from studies on telomeres, that G-quadruplexes form and can slow down or block DNA replication in vivo.[37,38] In *S. cerevisiae*, G4 structures slow down replication forks, a reaction that can be counteracted by the Pif1 helicase only when G4 are on the leading strand template.[23,24]

In contrast to *S. cerevisiae* (in which no correlation between Oris and G4 was observed), metazoan Oris display a G4 at 160 (Drosophila) (data not shown) and 280 bp (mouse) 5' from the replication initiation peaks. In mammalian cells, two DNA helicases, FANCJ and REV1, might operate together at the fork to facilitate the replication of a subset of G-quadruplex-forming sequences.[39] Interestingly, like in *S. cerevisiae*, the efficiency of replication of a plasmid, including the chicken β-*globin* Oris, can be strikingly reduced when a G4 sequence is placed on the leading-strand template but not on the lagging-strand template in *REV1* mutant cells.[40]

In view of these data, the presence of a G-quadruplex in 74% of mouse Oris and nearly 43% of Drosophila Oris (data not shown) specifically on the leading strand template suggests a new major role for G4 in the localization and/or in controlling the activation or repression of replication initiation in metazoans. A majority of mouse Oris associated with genes have an OGRE/G4 sequence, suggesting that this structure may also help coordinating replication initiation and transcription of the associated genes. This coordination was recently studied in viruses, where a CGC-motif upstream of Ori is bound by a multi-protein complex, which results in a physical barrier to block replication, allowing transcription of the adjacent genes.[41]

## Materials and Methods

**Characterization of OGRE occurrence.** Enriched motifs in Oris were identified using the MEME bioinformatics suite.[42] A fifth-order Markov model was generated as a background distribution model. From 5% of mouse and Drosophila Oris, 2 kb of DNA sequences centered on the NS peak were randomly selected. For each motif, an E-value was computed. E-values are commonly used for assigning significance to the optimal reported motifs. The most significant motif for mouse and Drosophila Oris was retrieved. Moreover, independent analyses were performed to show that the results were not dependent on the Ori sample. As an additional negative control, randomly selected genomic sequences were also analyzed. Genomic frequencies of OGRE motifs were computed using the FIMO software of the MEME suite.[42] Only occurrences with a q-value lower than 1% were used in this study. The q-value is the estimated false discovery rate if the occurrence is accepted as significant.[43]

**G-quadruplex (G4) elements prediction.** The G-quadruplex map was generated using the free Quadparser software[44] at http://www.quadruplex.org. Quadparser identifies putative quadruplexes in DNA sequences. The default G-quadruplex definition was used: $G3N_{1-7}G3N_{1-7}G3N_{1-7}G3$.

**Nucleotide skew computation.** The nucleotide skew was obtained by computing the following formula by 100 bp windows with a shift of 50 bp: (1) skew GC: (#G - #C)/(#G + #C); (2) skew TA: (#T - #A)/(#T + #A);[45] overall skew: skew GC + skew TA. Windows with a skew above 0.5 (in absolute value) were then selected to create the map of skewed regions.

**Computation of the overlap between Oris and genomic features.** For each profile (Drosophila and mouse cells), the overlap of at least 1 bp between Oris and different genomic features (OGRE occurrences, G4 elements and CGI) was computed. The reciprocal overlap was also performed. The associations are illustrated by Venn diagrams.

**Density of origins and other genome features.** Density analysis was performed to compare specific data distribution along chromosome 11 of the mouse genome. The coordinates of the specific regions (Oris and OGRE occurrence) were retrieved. Each nucleotide inside specific regions was flagged as 1 (if belonging to one specific region) and 0 (if not belonging to one specific region). One sliding window of 70 kb was used to compute the data frequency per window.[2]

**Nascent strand signal profile around specific features.** For Drosophila and mouse cells, the NS signal profile was studied around specific features: (1) OGRE occurrences; (2) OGRE occurrences overlapping or not with CGI;[45] G4 elements.

More precisely, for each profile (Drosophila and mouse cells), the middle of each specific feature was taken as "Local center" (Lcent). For each nucleotide position around every Lcent (Lcent - 5 kb to Lcent + 5 kb), the p-values were retrieved (see Sup. Experimental Procedures of Cayrou, 2011 #2755) and merged in a matrix (rows representing the nucleotide coordinate/position and columns representing Lcent). The strand was also considered. Thus, when the OGRE occurred on the minus strand, the nucleotide positions and the associated p-values were reversed. To obtain only one overall p-value distribution around the set of Lcent, the p-values were combined using a Chi-Square distribution.[46] To visualize the combined p-value distributions around specific features, results were plotted using the transformation "-log(p-value)" and labeled as "NS signal strength."

**Bidirectional/unidirectional oris relative to OGRE occurrences.** Three classes of Oris were defined according to the profile of the OGRE occurrences overlapping with the Ori sequence. The first class included Oris in which OGREs occurred mostly on the (+) strand, whereas the second class included Oris with OGRE occurring mostly on the (-) strand. The threshold to define the affiliation to the first or second class was 10: i.e., OGRE occurring 10 times more frequently in one direction than in the other direction. Oris affiliated to the first or second class were called unidirectional Oris, as the OGRE occurrence was mainly oriented in the same direction. The third class included Oris with OGRE occurring in both directions. These Oris were thus called bidirectional Oris.

**Intrinsic nucleosome occupancy prediction.** Intrinsic nucleosome occupancy was predicted using a previously published model trained on yeast in

vitro data.[36] The concentration and temperature parameters were taken to be 0.1 and 1, respectively.


### Acknowledgments

The research leading to these results has received funding from the European Research Council under the Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 233339". This work was also supported by the 'Agence Nationale de la Recherche' (ANR) (ANR-08-BLAN-0092-0), the ARC and the 'Ligue Nationale Contre le Cancer' (LNCC).


### References


1. Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, et al. Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc Natl Acad Sci USA 2008; 105:15837-42; PMID:18838675; http://dx.doi.org/10.1073/pnas.0805208105.
2. Cayrou C, Coulombe P, Vigneron A, Stanojcic S, Ganier O, Peiffer I, et al. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res 2011; 21:1438-49; PMID:21750104; http://dx.doi.org/10.1101/gr.121830.111.
3. Rajewska M, Wegrzyn K, Konieczny I. AT-rich region and repeated sequences—the essential elements of replication origins of bacterial replicons. FEMS Microbiol Rev 2011; Epub Ahead of print; PMID:22092310; http://dx.doi.org/10.1111/j.1574-6976.2011.00300.x.
4. Matsunaga F, Forterre P, Ishino Y, Myllykallio H. In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. Proc Natl Acad Sci USA 2001; 98:11152-7; PMID:11562464; http://dx.doi.org/10.1073/pnas.191387498.
5. Lundgren M, Andersson A, Chen L, Nilsson P, Bernander R. Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination. Proc Natl Acad Sci USA 2004; 101:7046-51; PMID:15107501; http://dx.doi.org/10.1073/pnas.0400656101.
6. Robinson NP, Dionne I, Lundgren M, Marsh VL, Bernander R, Bell SD. Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. Cell 2004; 116:25-38; PMID:14718164; http://dx.doi.org/10.1016/S0092-8674(03)01034-1.
7. McBride AA. Replication and partitioning of papillomavirus genomes. Adv Virus Res 2008; 72:155-205; PMID:19081491; http://dx.doi.org/10.1016/S0065-3527(08)00404-1.
8. Muylaert I, Tang KW, Elias P. Replication and recombination of herpes simplex virus DNA. J Biol Chem 2011; 286:15619-24; PMID:21362621; http://dx.doi.org/10.1074/jbc.R111.233981.
9. Murakami Y, Ito Y. Transcription factors in DNA replication. Front Biosci 1999; 4:824-33; PMID:10577391; http://dx.doi.org/10.2741/Murakami.
10. Dellarole M, Sánchez IE, de Prat Gay G. Thermodynamics of cooperative DNA recognition at a replication origin and transcription regulatory site. Biochemistry 2010; 49:10277-86; PMID:21047141; http://dx.doi.org/10.1021/bi1014908.
11. Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM. Conserved nucleosome positioning defines replication origins. Genes Dev 2010; 24:748-53; PMID:20351051; http://dx.doi.org/10.1101/gad.1913210.
12. Breier AM, Chatterji S, Cozzarelli NR. Prediction of *Saccharomyces cerevisiae* replication origins. Genome Biol 2004; 5:22; PMID:15059255; http://dx.doi.org/10.1186/gb-2004-5-4-r22.
13. Maundrell K, Hutchison A, Shall S. Sequence analysis of ARS elements in fission yeast. EMBO J 1988; 7:2203-9; PMID:3046932.
14. Okuno Y, Satoh H, Sekiguchi M, Masukata H. Clustered adenine/thymine stretches are essential for function of a fission yeast replication origin. Mol Cell Biol 1999; 19:6699-709; PMID:10490609.
15. Cotobal C, Segurado M, Antequera F. Structural diversity and dynamics of genomic replication origins in *Schizosaccharomyces pombe*. EMBO J 2010; 29:934-42; PMID:20094030; http://dx.doi.org/10.1038/emboj.2009.411.
16. Zellner E, Herrmann T, Schulz C, Grummt F. Site-specific interaction of the murine pre-replicative complex with origin DNA: assembly and disassembly during cell cycle transit and differentiation. Nucleic Acids Res 2007; 35:6701-13; PMID:17916579; http://dx.doi.org/10.1093/nar/gkm555.
17. Ghosh M, Kemp M, Liu G, Ritzi M, Schepers A, Leffak M. Differential binding of replication proteins across the human c-myc replicator. Mol Cell Biol 2006; 26:5270-83; PMID:16809765; http://dx.doi.org/10.1128/MCB.02137-05.
18. Abdurashidova G, Deganuto M, Klima R, Riva S, Biamonti G, Giacca M, et al. Start sites of bidirectional DNA synthesis at the human lamin B2 origin. Science 2000; 287:2023-6; PMID:10720330; http://dx.doi.org/10.1126/science.287.5460.2023.
19. Dijkwel PA, Hamlin JL. The Chinese hamster dihydrofolate reductase origin consists of multiple potential nascent-strand start sites. Mol Cell Biol 1995; 15:3023-31; PMID:7760799.
20. Maizels N. Dynamic roles for G4 DNA in the biology of eukaryotic cells. Nat Struct Mol Biol 2006; 13:1055-9; PMID:17146462; http://dx.doi.org/10.1038/nsmb1171.
21. Wong HM, Stegle O, Rodgers S, Huppert JL. A toolbox for predicting g-quadruplex formation and stability. J Nucleic Acids 2010; 2010; PMID:20725630; http://dx.doi.org/10.4061/2010/564946.
22. Cech TR. Beginning to understand the end of the chromosome. Cell 2004; 116:273-9; PMID:14744437; http://dx.doi.org/10.1016/S0092-8674(04)00038-8.
23. Lopes J, Piazza A, Bermejo R, Kriegsman B, Colosio A, Teulade-Fichou MP, et al. G-quadruplex-induced instability during leading-strand replication. EMBO J 2011; 30:4033-46; PMID:21873979; http://dx.doi.org/10.1038/emboj.2011.316.
24. Paeschke K, Capra JA, Zakian VA. DNA replication through G-quadruplex motifs is promoted by the *Saccharomyces cerevisiae* Pif1 DNA helicase. Cell 2011; 145:678-91; PMID:21620135; http://dx.doi.org/10.1016/j.cell.2011.04.015.
25. Siddiqui-Jain A, Grand CL, Bearss DJ, Hurley LH. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proc Natl Acad Sci USA 2002; 99:11593-8; PMID:12195017; http://dx.doi.org/10.1073/pnas.182256799.
26. Huppert JL, Balasubramanian S. G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res 2007; 35:406-13; PMID:17169996; http://dx.doi.org/10.1093/nar/gkl1057.
27. Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, et al. Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol 2008; 6:245; PMID:18842067; http://dx.doi.org/10.1371/journal.pbio.0060245.
28. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 2009; 458:362-6; PMID:19092803; http://dx.doi.org/10.1038/nature07667.
29. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, et al. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol 2009; 16:847-52; PMID:19620965; http://dx.doi.org/10.1038/nsmb.1636.
30. Grainge I, Gaudier M, Schuwirth BS, Westcott SL, Sandall J, Atanassova N, et al. Biochemical analysis of a DNA replication origin in the archaeon *Aeropyrum pernix*. J Mol Biol 2006; 363:355-69; PMID:16978641; http://dx.doi.org/10.1016/j.jmb.2006.07.076.
31. Gaczynska M, Osmulski PA, Jiang Y, Lee JK, Bermudez V, Hurwitz J. Atomic force microscopic analysis of the binding of the *Schizosaccharomyces pombe* origin recognition complex and the spOrc4 protein with origin DNA. Proc Natl Acad Sci USA 2004; 101:17952-7; PMID:15598736; http://dx.doi.org/10.1073/pnas.0408369102.
32. Remus D, Beall EL, Botchan MR. DNA topology, not DNA sequence, is a critical determinant for Drosophila ORC-DNA binding. EMBO J 2004; 23:897-907; PMID:14765124; http://dx.doi.org/10.1038/sj.emboj.7600077.
33. Dueber EC, Costa A, Corn JE, Bell SD, Berger JM. Molecular determinants of origin discrimination by Orc1 initiators in archaea. Nucleic Acids Res 2011; 39:3621-31; PMID:21227921; http://dx.doi.org/10.1093/nar/gkq1308.
34. Sun D, Guo K, Shin YJ. Evidence of the formation of G-quadruplex structures in the promoter region of the human vascular endothelial growth factor gene. Nucleic Acids Res 2011; 39:1256-65; PMID:20959293; http://dx.doi.org/10.1093/nar/gkq926.
35. Tillo D, Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Field Y, et al. High nucleosome occupancy is encoded at human regulatory sequences. PLoS One 2010; 5:9129; PMID:20161746; http://dx.doi.org/10.1371/journal.pone.0009129.
36. Tillo D, Hughes TRG. G + C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 2009; 10:442; PMID:20028554; http://dx.doi.org/10.1186/1471-2105-10-442.
37. Schaffitzel C, Berger I, Postberg J, Hanes J, Lipps HJ, Plückthun A. In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. Proc Natl Acad Sci USA 2001; 98:8572-7; PMID:11438689; http://dx.doi.org/10.1073/pnas.141229498.
38. Sfeir A, Kosiyatrakul ST, Hockemeyer D, MacRae SL, Karlseder J, Schildkraut CL, et al. Mammalian telomeres resemble fragile sites and require TRF1 for efficient replication. Cell 2009; 138:90-103; PMID:19596237; http://dx.doi.org/10.1016/j.cell.2009.06.021.
39. Sarkies P, Murat P, Phillips LG, Patel KJ, Balasubramanian S, Sale JE. FANCJ coordinates two pathways that maintain epigenetic stability at G-quadruplex DNA. Nucleic Acids Res 2011; Epub Ahead of Print; PMID:22021381; http://dx.doi.org/10.1093/nar/gkr868.
40. Sarkies P, Reams C, Simpson LJ, Sale JE. Epigenetic instability due to defective replication of structured DNA. Mol Cell 2010; 40:703-13; PMID:21145480; http://dx.doi.org/10.1016/j.molcel.2010.11.009.
41. Khalil MI, Arvin A, Jones J, Ruyechan WT. A sequence within the varicella-zoster virus (VZV) OriS is a negative regulator of DNA replication and is bound by a protein complex containing the VZV ORF29 protein. J Virol 2011; 85:12188-200; PMID:21937644; http://dx.doi.org/10.1128/JVI.05501-11.

42. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 2009; 37:202-8; PMID:19458158; http://dx.doi.org/10.1093/nar/gkp335.
43. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003; 100:9440-5; PMID:12883005; http://dx.doi.org/10.1073/pnas.1530509100.
44. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res 2005; 33:2908-16; PMID:15914667; http://dx.doi.org/10.1093/nar/gki609.
45. Sahyoun N, Wolf M, Besterman J, Hsieh TS, Sander M, LeVine H, 3rd, et al. Protein kinase C phosphorylates topoisomerase II: topoisomerase activation and its possible role in phorbol ester-induced differentiation of HL-60 cells. Proc Natl Acad Sci USA 1986; 83:1603-7; PMID:3006058; http://dx.doi.org/10.1073/pnas.83.6.1603.
46. Fisher RA. Statistical Methods for Research Workers. Oliver and Boyd 1932; London.